Chatbots and Mental Health: Insights into the Safety of Generative AI

Year: 2023 | Citations: 103 | Authors: Julian De Freitas, A. U■uralp, Zeliha O■uz■U■uralp, Stefano Puntoni

Abstract

Chatbots are now able to engage in sophisticated conversations with consumers. Due to the ‘black box’ nature of the algorithms, it is impossible to predict in advance how these conversations will unfold. Behavioral research provides little insight into potential safety issues emerging from the current rapid deployment of this technology at scale. We begin to address this urgent question by focusing on the context of mental health and “companion AI”: applications designed to provide consumers with synthetic interaction partners. Studies 1a and 1b present field evidence: actual consumer interactions with two different companion AIs. Study 2 reports an extensive performance test of several commercially available companion AIs. Study 3 is an experiment testing consumer reaction to risky and unhelpful chatbot responses. The findings show that (1) mental health crises are apparent in a non■negligible minority of conversations with users; (2) companion AIs are often unable to recognize, and respond appropriately to, signs of distress; and (3) consumers display negative reactions to unhelpful and risky chatbot responses, highlighting emerging reputational risks for generative AI companies.